

SHOULD YOU TRUST REDDIT WITH YOUR MONEY?

A MACHINE LEARNING APPROACH TO ESTIMATING STOCK
RETURNS

LOISE HEDBERG

OLLE RIBBERHEIM

Bachelor Thesis

Stockholm School of Economics

2021



Should You Trust Reddit with Your Money? — A Machine Learning Approach to Estimating Stock Returns

Abstract:

We replicate the results of *Empirical Asset Pricing via Machine Learning* by Gu, Kelly, and Xiu (2020), and make an extension by using a daily interval instead of a monthly. Furthermore, we make an extension of the topic by generating data from wallstreetbets, in order to investigate whether there is explanatory value of the subreddit to predict stock returns. The activity and sentiment attributable to different companies on wallstreetbets are analyzed and used in the prediction. Boosted decision trees and artificial neural networks are modelled and evaluated based on root mean square error and R^2 and compared between same day returns and next day returns, as well as with and without the reddit data in order to determine the explanatory value of it. We find some explanatory value of the wallstreetbets posts for the same day returns, however no predictive value for the next day returns, in line with the Efficient market-hypothesis.

Keywords:

Empirical asset pricing, Machine learning, Social media, Market efficiency, Stock returns

Authors:

Loise Hedberg (24359)

Olle Ribberheim (24304)

Tutors:

Adrien d'Avernas, Assistant Professor, Department of Finance

Karl Wärneryd, Associate Professor, Department of Economics

Examiner:

Adrien d'Avernas, Assistant Professor, Department of Finance

Bachelor Thesis

Bachelor Program in Business and Economics

Stockholm School of Economics

© Loise Hedberg and Olle Ribberheim, 2021

1 Introduction

Asset pricing is the act of estimating the risk premium of an asset, and can be estimated by using various data of a stock. The risk premium of an asset is the surplus yield of the asset above the risk-free rate, and acts as a compensation for the risk the investor is taking when investing in a stock compared to the risk free rate. The risk premium is important when valuing financial assets. The stock market can be extremely lucrative for investors, if navigated correctly. Risk premiums are however hard to measure, since unpredictable shocks from the market affect them and make it problematic to distinguish the pure value of the asset from outside influence.

The subforum wallstreetbets on the social news website reddit caused the long-time poorly performing stock of Gamestop to skyrocket in January 2021. (TradeNews.com, 2021) Before, hedge funds such as Melvin Capital started to short sell the Gamestop stocks, lending stocks and selling them, in hope that their price would decline even more, so that they could be rebought for a lesser price and the hedge fund would profit from the difference. In reaction to the short selling, the people on wallstreetbets started propagating for purchasing large volumes of the stock, causing the stock to increase its value. The value increase of the stock forced the hedge funds to buy back the stock to minimize their loss while the underwriters of options on gamestop had to buy stocks to delta hedge their positions, known as a “gamma squeeze”. This in turn caused the value of the stock to increase even more. The stock reached its peak on January 27th 2021, with an adjusted closing price of \$347,51, having increased by 1553% compared to its value on 28th December, 2020 of \$21. (Yahoo Finance, 2021) Simultaneously, Melvin Capital lost roughly \$4.5 billions during January 2021, largely because of this, going from assets of \$12.5 billions to assets of \$8 billions. (BusinessInsider.com, 2021) This event revealed the potential power of individual investors. Individual investors, also known as retail investors, are a large segment of securities trading, with a significant impact on the stock market and the sentiment of it. (Investopedia.com, 2021) Retail investors today make up 20% of equity volume trades in the USA, for example. (Financial Times, 2021) We seek to investigate to which extent the sentiment analysis of the posts and comments on wallstreetbets can be used to determine and foresee the risk premium of stocks, and whether the data from the social news website could be used in addition to other features to improve the measurement of equity risk premiums.

We seek to partially replicate Gu, Kelly, and Xiu (2020) by contributing to a machine learning toolbox in Python which estimates the equity risk premiums of stocks. We also add the extension of using daily data instead of, as used by Gu, Kelly, and Xiu (2020), monthly data from the Compustat database, provided by S&P global market intelligence. In our replication, we use a selection of the features deemed most important in the paper, of which the most important groups are recent price trends, liquidity variables,

and volatility variables. Some of the features included in our toolbox are short-term reversal, stock momentum, turnover, total and idiosyncratic return volatility, earnings-to-price, and asset growth. Our replication investigates if using daily, instead of monthly, data improves the measurement of the risk premiums as done by Gu, Kelly, and Xiu (2020).

We seek to contribute to the literature twofold, firstly we aim to contribute by extracting and analyzing posts and comments from wallstreetbets through neural networks and boosted decision trees to inquire if this sentiment and API feature has a predictive value for the stocks of the companies mentioned in the posts. Secondly, by replicating Gu, Kelly, and Xiu (2020) with modifications, with the extension of using daily data instead of monthly to investigate whether using more frequent data improves the measurements of risk premiums. The authors find that using monthly data yielded better predictions than when using annual data, and it is of interest to further see if daily data yields better results than monthly data.

We use data from the time period 1980 to 2019 in our replication of the toolbox by Gu, Kelly, and Xiu (2020). The data is from Center for Research in Security Prices (CRSP), which has been accessed through Wharton Research Data Services (WRDS).

For our analysis with wallstreetbets, we built and used a script to scrape the comments and posts from the subreddit from its beginning in 2012 to the end of 2020. We then analyzed them and divided them into groups of sentiments connected to companies, which we then analyzed through boosted decision trees and neural networks.

Our results show that there is explanatory value in the wallstreetbets posts for the same day stock returns, but no predictive value for foreseeing the results of the next day. The replication of Gu, Kelly, and Xiu (2020) succeeded on monthly data, having a considerable higher return than the benchmark of equally weighted positions in each stock, whereas the extension with daily data had low explanatory value, and performed close to the equally weighted portfolio.

This paper will be presented in this order: In section 2, our literature review, in section 3 the methodology used is described, in section 4, the description of the data is presented. In section 5, the empirical results, in section 6, our discussion and finally our conclusion in section 7.

2 Literature Review

The first significant model of asset pricing, *The Capital Asset Pricing Model* (CAPM), was introduced by Sharpe (1964), Lintner (1965), and Mossin (1966) and is still widely used. The CAPM obtains the expected rate of return from the risk-free rate, the beta, and the market risk premium. Since the 1960s, substantial research has been performed on asset pricing, and produced improved and more advanced models for asset pricing.

Empirical Asset Pricing via Machine Learning by Gu, Kelly, and Xiu (2020) analyzes machine learning methods for valuing assets. The study shows that machine learning algorithms, especially neural networks and trees, can be useful for predicting and measuring equity risk premiums, and prove more successful than popular regression methods. Furthermore, the authors identify the most successful predictor variables for asset pricing: variations of liquidity, momentum, and volatility. We replicate this paper by using the most important predictor variables, but we use daily data instead of monthly data. It is of interest to see whether this improves the prediction, since Gu, Kelly, and Xiu find that monthly data improved the prediction compared to yearly data. However, it is possible that using daily data might include more noise and interfere more with our predictions, making it worse than using monthly data.

Tumarkin and Whitelaw (2001) collect data from the online forum *RagingBull.com*, which analyzes finance. The authors analyze frequency and sentiment activity for companies on the forum together with abnormal stock trading volumes and stock prices, through regressions. It is concluded that investor opinion changes are correlated with unusual industry-adjusted returns and unusually high trading volumes on days with unusually high post activity in the forum. Tumarkin and Whitelaw find no evidence of forum posts affecting the stock returns or trading volume, however they suspect a reverse causality between the two, which is in accordance with market efficiency. It is of interest to see whether our study as well produces results which are in line with the Efficient-market hypothesis, as defined by Fama et al (1969). According to this hypothesis, the price of the stock always represents all the information that is available. Therefore, abnormal returns would be unattainable since returns would not be affected by for example activity on internet forums.

Yang, Zhu, Cheng (2020) perform a similar investigation, and find results that indicate that prediction of trading volumes is possible through analyzing the activity on a stock forum, which, on the contrary to Tumarkin and Whitelaw (2001), is not in line with the Efficient Market-hypothesis. Yang, Zhu, Cheng (2020) use data from China's largest stock message board (Eastmoney.com), and create a "Sentiment index" and a "Panic index" to represent the emotions of investors on the forum. Yang, Zhu, and Cheng (2020), unlike Tumarkin and Whitelaw (2001), find through regression models that you can predict

trading volumes through analyzing the board; that a high “Sentiment Index” of a stock causes abnormal trading volumes, and that the “Panic Index” could help foresee a stock crash.

We investigate whether the activity, analyzed by sentiment and frequency, of the most popular companies on WSB has any predictive value for the next day’s stock return, as well for the same day. Our method of analyzing the internet posts includes sentiment and occurrence analysis of an online stock forum, in similarity to Tumarkin and Whitelaw (2001) and Yang, Zhu, Cheng (2020), but differs from them in both the forum which is analyzed and the type of analysis used. We use neural networks and trees instead of regressions and event studies. Since Tumarkin and Whitelaw (2001) and Yang, Zhu, Cheng (2020) get results which differ widely in however one can foresee stock behaviour through analysis of internet postings, we seek to explore how our results compare to these and the Efficient market-hypothesis.

John Maynard Keynes wrote about the concept of the Keynesian beauty contest in his book *The General Theory of Employment, Interest and Money* (1936). The concept, which contrasts the Efficient market-theory, compares a beauty contest, where participants should choose the most popular faces from a number of photographs, to the stock market. The similarity is that a more advanced strategy would be to not choose the faces/stocks that you believe are the most attractive, but the ones that you believe that the other participants believe will be the most popular ones. A parallel can be drawn to our exploration of how WSB affects the stock market where, if they turn out to have an effect on the pricing of stocks, it could be seen as an indication of the notion of the Keynesian beauty contest. If the concept holds true, participants on WSB would choose stocks that they believe are popular among the public and, as an example, a large amount of positive posts about a certain stock could then signal that the stock is popular, and hence convince others to buy it.

3 Methodology

A library to replicate Gu, Kelly, and Xiu (2020) is developed together with PhD students and students pursuing a master’s degree at SSE. The library consists of two parts. Firstly, a script which connects to the CRSP and Compustat databases and generates a dataset similar to the one described in Gu, Kelly, and Xiu, but with daily data instead of monthly, from 1980 to 2020. The features with the highest explanatory value for the model from Gu, Kelly, and Xiu are defined and cleaned. Secondly, a neural network which is trained on the data with the purpose of predicting stock movements. For each day, it chooses the ten stocks which it predicts will perform the best and takes long positions in those.

To compile the reddit data, the API PRAW (The Python Reddit API Wrapper) is used to extract all comments and posts published on wallstreetbets from the first of January 2012 to the last of December 2020. Since the server extracts all posts and comments the moment they are posted, the data set reflects the subreddit from the historic day, even if posts or comments are deleted retrospectively. However, the number of upvotes, the hierarchy of submissions and comments, and occurrence on hot pages and r/all is not captured with this methodology. Although it is true that the data does not fully represent the state of the subreddit at the current time being, it is more consequent and captures all published posts and activity.

After all comments and submissions are downloaded, a dictionary, seen in appendix 1 in section 8.1, which links names and tickers associated with certain companies to their respective company, is used to parse through all comments and submissions and save down all comments and submissions which contain the given ticker or name of the current day. For each day, the frequency of comments and submissions where each word occurred at least once separately in is calculated and each of those comments is saved down. A sentiment for each stock is determined for each day by using DistilBERT's sentiment analysis transformer Hugging Face to determine the sentiment in each comment and submission where the name or ticker of a company occurs. Hugging Face takes a string of text as input and returns a value between -1 and 1, where -1 is extremely negative and 1 extremely positive. A mean for each day is calculated for each stock based on the comments and submissions where the linked names occur.

A normalized frequency measure and an average sentiment for each company and date is calculated for further analysis and modelling. The frequency measure, defined in equation 1, is the number of occurrences of words corresponding to a given company, x_i over the number of occurrences of words corresponding to any company for that given day, x . We sort all companies by the sum of frequency measures to find the most relatively mentioned companies and look more into the top 100. Companies which are not mentioned especially often become hard to analyze and predict the stock performance for since there are small amounts of data and large time gaps between it being mentioned. Among the top 100 mentioned companies, we removed the companies which were at high risk of not being intended in the sentence, where the company had a ticker which is equal to a common word. For example, the company "World Health Energy Holdings, Inc." has the ticker "WHEN", which is also a commonly used English and not only a ticker. However, we choose not to remove the companies Facebook and Twitter since, even though people may mention them and not intend to comment on the stock of these companies, they are still referring to them as companies unlike words that are just common, like "when" and "we".

$$\text{Frequency}_i = \frac{\sum_{i=1}^N x_i}{\sum_{i=1}^N x} \quad (1)$$

Thereafter, to determine the explanatory value of wallstreetbets for stock returns, both a boosted decision tree and a neural network are used to predict the stock return of the stocks. Firstly, the models are created solely based on the adjusted closing returns of the S&P 500 index adjusted closing returns. Secondly, they are created based on both the adjusted closing returns of the S&P 500 and the sentiment and frequency data corresponding to the specific stock and date. Additionally, the S&P 500 data set is shifted one date so that the models try to predict the stock returns the following day instead of the same day, and identical boosted decision trees and neural networks are trained on this new data set. The data sets are split into a training data set and a test data set, where the training data set consists of 90% of the data, randomly chosen, and the test data set consists of the remaining 10%. By creating the model based on the training data and assessing it based on the test data, we can find overfitted and unsuitable models since the model does not have seen the data in the test set previously. By comparing the accuracy of the models, it is possible to determine and quantify the explanatory value of the reddit data to determine stock returns.

When implementing the boosted decision tree, the XGBoost library for Python is used, and more explicitly the XGB class. Squared error is used as the objective function and a depth of 50 layers is used when implementing the boosted decision tree, which has proven to be an efficient trade-off between accuracy and stability. For the evaluation method, the root mean square log error (RMSLE) is used, which is the default option and similar to RMSE in terms of characteristics. The model is trained on the training data before evaluating it on the test data.

The neural network is implemented in Python using the TensorFlow library. The rectified linear unit (relu) activation is used and we are using normal distributed tensors as kernel initializers. Furthermore, adaptive moment estimation (adam) is used as an optimizer and mean square error (MSE) and mean absolute error (MAE) as loss evaluation metric. We use a batch size of 100 and 150 epochs for the training on the training data set, and then evaluate it on the test data set.

The accuracy of the models is determined by calculating RMSE and R^2 . RMSE, root mean square error, is defined as

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}. \quad (2)$$

RMSE measures accuracy of the model and is suitable for quantifying the error when we want to penalize large errors since it takes the square of the error instead of the absolute value, which the mean absolute error (MAE) does. The smaller the RMSE, the more

accurate the model.

R^2 is another accuracy measure and can be interpreted as how much of the deviance in the response variable that can be explained by deviances in the explanatory variable. That is

$$\begin{aligned}\bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i, \\ \text{SS}_{\text{tot}} &= \sum_i (y_i - \bar{y})^2, \\ \text{SS}_{\text{res}} &= \sum_i (y_i - f_i)^2 = \sum_i e_i^2, \\ R^2 &= 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}}.\end{aligned}\tag{3}$$

R^2 ranges between 0 and 1, where values close to 1 indicate high explanatory value while close to 0 indicates no or low explanatory value. Adjusted R^2 is a similar measure but penalizes more features and is given by:

$$\text{Adjusted}R^2 = 1 - \frac{(1 - R^2) \cdot (n - 1)}{(n - k - 1)}\tag{4}$$

where n is the number of data points and k is the number of features. Observe that Adjusted R^2 can take negative values, whereas R^2 only can take non-negative values.

4 Data Description

For the replication of Gu, Kelly, and Xiu (2020), we collect and use stock data from the database of Center for Research in Security Prices (CRSP), which is accessed through Wharton Research Data Services (WRDS). We gather daily stock market data from CRSP and merge it with quarterly and annual data from WRDS, where each data point from CRSP is merged with the most recently published quarterly and annual data, to the highest extent possible in order to replicate the scenario for each day with the data available at that point in time. Examples of the data with the imported and calculated features can be found in appendix 8.2 , as well as the definitions of the features.

For our wallstreetbets extension, we build and use a reddit scraper script in Python to scrape all comments and submissions from the subreddit wallstreetbets from its creation date in January 31 2012 to December 31 2020. The number of posts amounted to 2982474. After collecting the data in the form of comments and submissions with a Python script, we use a dictionary which links names and tickers corresponding to a specific company, for example, the words "Tesla" and "TSLA" are linked to Tesla, Inc. The regex which the parser is based on looks for independent words in the comment/submission. For example, for the company "Advanced Micronic Devices" with the ticker "AMD", the parser does not find posts with words such as "ramdrive" or "amdy", even though these words contain "amd", since the word is not independent. The number of companies amounted to 7449. A Python script is used to analyze the frequency of each company, and we then use the 22 most commonly mentioned companies on WSB for our further analysis.

Analyzing the data from wallstreetbets delimits the analysis to what the users of the forum decide to post online. The population of wallstreetbets may not be representative of the aggregate stock buyers and sellers. They amount to 10.1 million users as of May 10th 2021 (*wallstreetbets.com*, 2021), and are described as “mostly young, mostly male amateur investors” on the website Forbes. (Forbes.com, 2021) In other words, they may not be representative for the population of all stock traders in the world. As a result, their sentiment and analysis of stocks may not coincide with that of the average stock trader, especially for “meme stocks” which are popular on the subreddit. “Meme stocks” are described as stocks that gain popularity and volume from an online hype, from for example reddit, instead of popularity which is based solely on the financial performance of a company. (TheBalance.com, 2021)

5 Empirical Results

5.1 Empirical Asset Pricing Replication

The results from the replication of Gu, Kelly, and Xiu (2020) are shown in the figures below. In figure 1, the correlation plot for different forecast periods are shown between the actual outcomes of the test data set and the predicted value. In figure 2, the correlation between the predicted values of the validation data set and the actual values are shown. In figure 3, the performance of the selected 10 monthly purchased stocks compared to the equally weighted portfolio of all stocks is shown. And lastly, in figure 4, the performance of the selected 10 daily purchased stocks compared to the equally weighted portfolio of all stocks is shown.

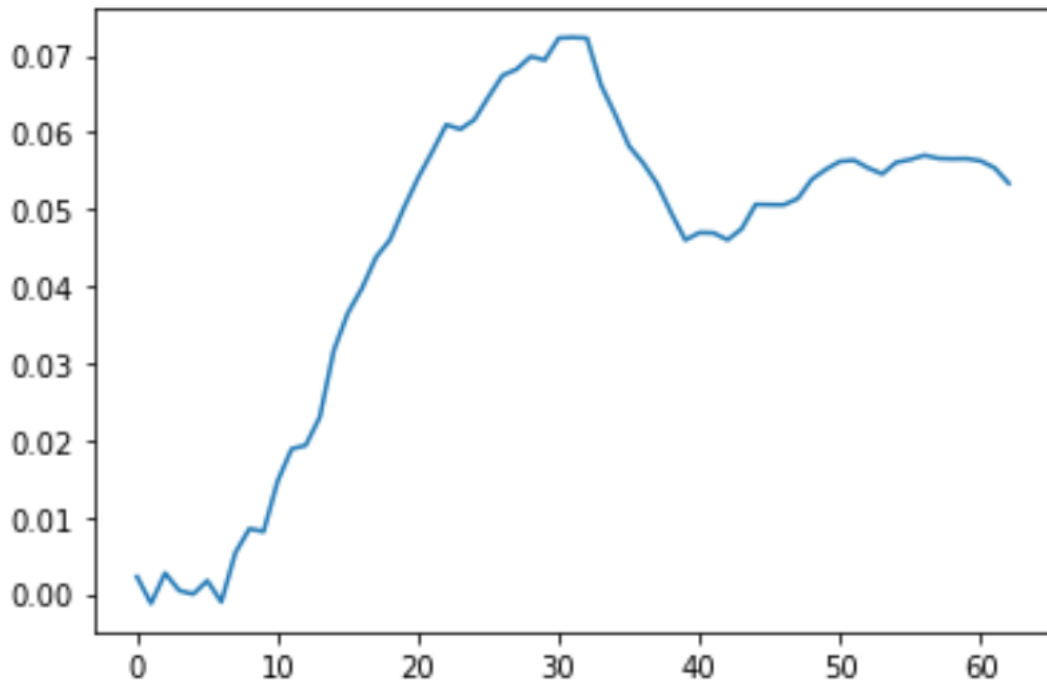


Figure 1: Correlation in the test dataset, between the predicted and true returns for various forecast horizons.

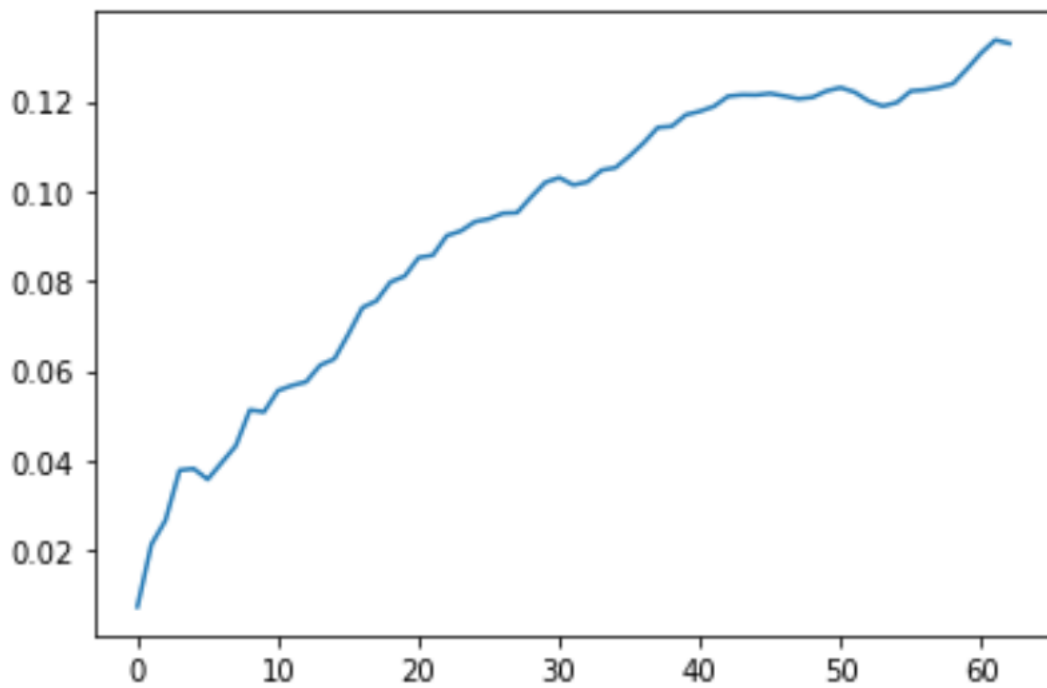


Figure 2: Correlation in the validation dataset, between the predicted and true returns for various forecast horizons.

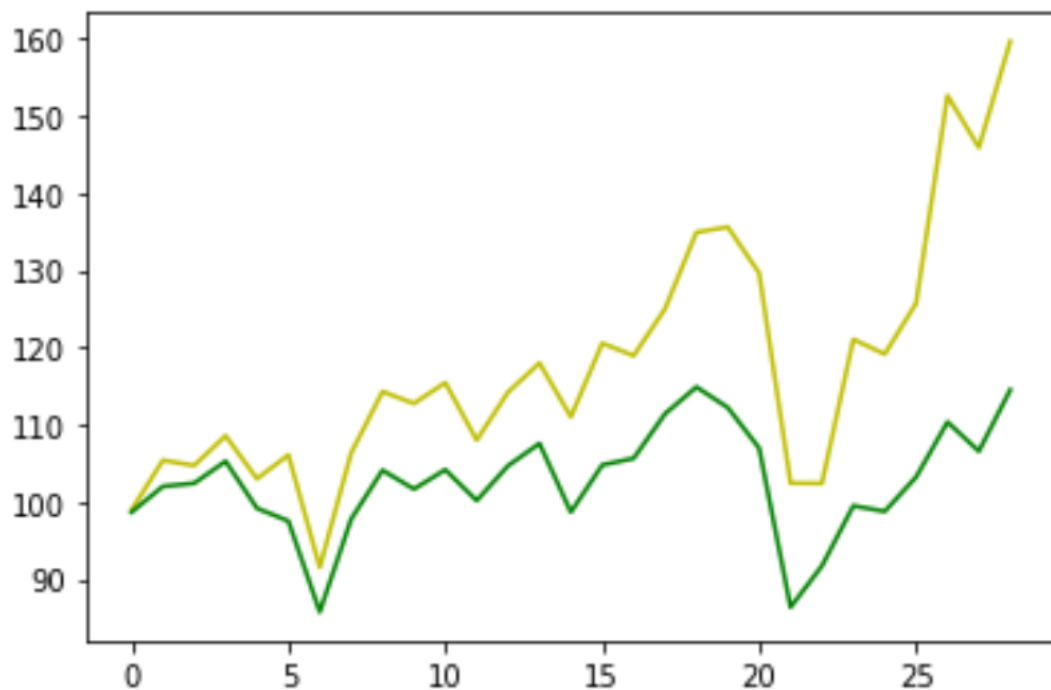


Figure 3: Portfolio development with monthly purchase of the top 10. The top 10 stocks are bought each month and then sold the next (yellow) vs an equal weight-portfolio (green). The graph is from von Essen and Wu, 2021.

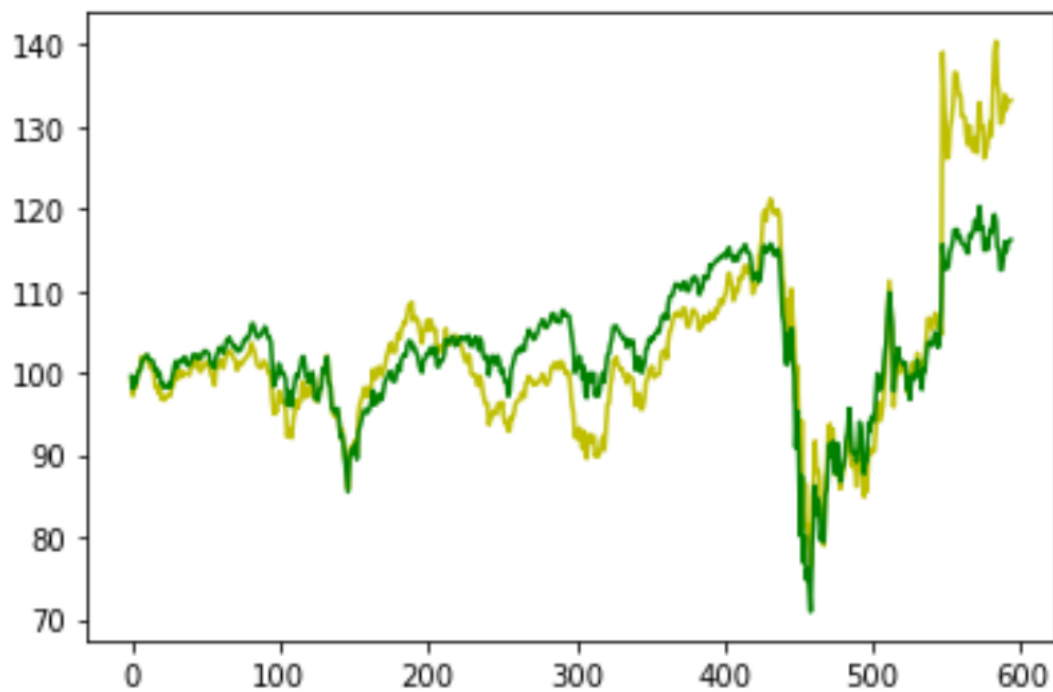


Figure 4: Portfolio development with daily purchase of the top 10. The top 10 stocks are bought each day and then sold the next (yellow) vs an equal weight-portfolio (green). The graph is from von Essen and Wu, 2021.

5.2 Wallstreetbets Sentiment Analysis

Among the 100 companies with the highest sum of relative frequencies, there were 22 companies which were intended. For the results, these 22 companies are ordered from highest frequency to lowest. That is, Apple Inc. has the highest frequency and Goldman Sachs Group Inc. has the lowest frequency out of the companies which we are analyzing.

A summary of the results from the boosted decision tree and the neural network is found in table 1 and table 2, whereas the complete data for each individual company is presented in table 3-10. In figure 5, 6, and 7, are scatter plots over the estimated and actual returns plotted for boosted decision trees based solely on the reddit data.

SAME DAY	Without WSB data	With WSB data
RMSE	3.782	3.595
R^2	0.095	0.140
adjusted R^2	0.089	0.086
NEXT DAY		
RMSE	4.245	4.017
R^2	0.011	0.031
adjusted R^2	0.004	-0.029

Table 1: Arithmetic mean of RMSE, R^2 , and adjusted R^2 for the 22 companies analyzed through boosted decision trees, on the same and the next day returns, and with and without wallstreetbets data.

SAME DAY	Without WSB data	With WSB data
RMSE	4.651	4.774
R^2	0.193	0.169
adjusted R^2	0.188	0.119
NEXT DAY		
RMSE	4.754	4.848
R^2	0.023	0.033
adjusted R^2	0.017	-0.025

Table 2: Arithmetic mean of RMSE, R^2 , and adjusted R^2 for the 22 companies analyzed through neural networks, on the same and the next day returns, and with and without wallstreetbets data.

Company	RMSE	R^2	adjusted R^2
Apple Inc.	1.809	0.244	0.240
Advanced Micro Devices Inc.	4.881	0.044	0.038
Tesla Inc.	4.494	0.035	0.029
Amazon.com Inc.	2.234	0.126	0.121
Facebook Inc.	2.464	0.132	0.127
Micron Technology Inc.	3.337	0.125	0.120
Netflix Inc.	3.592	0.044	0.038
Alibaba Group Holding Limited	2.549	0.078	0.072
JP Morgan Chase & Co.	1.588	0.340	0.337
Snap Inc.	5.637	0.010	0.000
Alphabet Inc.	5.394	0.035	0.025
Square Inc.	1.612	0.264	0.260
Twitter Inc.	4.070	0.095	0.088
Activision Blizzard Inc.	4.494	0.039	0.033
Intel Corporation	2.546	0.068	0.063
Gamestop Corporation	1.940	0.199	0.195
Beyond Meat Inc.	5.162	0.015	0.009
Palantir Technologies Inc. Class A	7.877	0.019	-0.006
Walmart Inc.	1.655	0.050	0.050
Virgin Galactic Holdings Inc.	9.053	0.006	0.001
Morgan Stanley	1.709	0.053	0.048
Goldman Sachs Group Inc.	5.106	0.075	0.063

Table 3: Boosted decision tree, same day returns, without WSB data.

Company	RMSE	R^2	adjusted R^2
Apple Inc.	1.655	0.337	0.321
Advanced Micro Devices Inc.	4.581	0.064	0.042
Tesla Inc.	4.287	0.050	0.029
Amazon.com Inc.	1.894	0.232	0.208
Facebook Inc.	2.391	0.146	0.124
Micron Technology Inc.	3.045	0.195	0.171
Netflix Inc.	3.083	0.067	0.035
Alibaba Group Holding Limited	2.306	0.133	0.103
JP Morgan Chase & Co.	2.196	0.423	0.359
Snap Inc.	5.418	0.008	-0.028
Alphabet Inc.	5.512	0.048	0.002
Square Inc.	1.384	0.448	0.429
Twitter Inc.	3.633	0.124	0.085
Activision Blizzard Inc.	3.618	0.175	0.125
Intel Corporation	2.529	0.105	0.035
Gamestop Corporation	2.412	0.182	-0.168
Beyond Meat Inc.	5.875	0.040	-0.002
Palantir Technologies Inc. Class A	6.401	0.040	-0.075
Walmart Inc.	1.742	0.090	0.053
Virgin Galactic Holdings Inc.	8.108	0.011	-0.027
Morgan Stanley	1.815	0.108	0.070
Goldman Sachs Group Inc.	5.206	0.053	0.001

Table 4: Boosted decision tree, same day returns, with WSB data.

Company	RMSE	R^2	adjusted R^2
Apple Inc.	2.534	0.006	0.001
Advanced Micro Devices Inc.	5.344	0.006	0.000
Tesla Inc.	5.048	0.010	0.004
Amazon.com Inc.	2.773	0.005	-0.001
Facebook Inc.	3.118	0.005	-0.001
Micron Technology Inc.	4.201	0.005	-0.001
Netflix Inc.	3.882	0.008	0.002
Alibaba Group Holding Limited	3.008	0.007	0.000
JP Morgan Chase & Co.	2.477	0.014	0.008
Snap Inc.	5.594	0.013	0.002
Alphabet Inc.	5.702	0.009	-0.001
Square Inc.	2.171	0.012	0.006
Twitter Inc.	4.808	0.010	0.002
Activision Blizzard Inc.	4.869	0.004	-0.002
Intel Corporation	2.832	0.005	-0.001
Gamestop Corporation	2.558	0.025	0.020
Beyond Meat Inc.	5.654	0.004	-0.002
Palantir Technologies Inc. Class A	8.009	0.022	-0.002
Walmart Inc.	1.839	0.030	0.024
Virgin Galactic Holdings Inc.	9.570	0.003	-0.002
Morgan Stanley	1.845	0.027	0.022
Goldman Sachs Group Inc.	5.546	0.011	-0.002

Table 5: Boosted decision tree, next day returns, without WSB data.

Company	RMSE	R^2	adjusted R^2
Apple Inc.	2.376	0.021	-0.003
Advanced Micro Devices Inc.	5.016	0.007	-0.016
Tesla Inc.	4.689	0.010	-0.013
Amazon.com Inc.	2.423	0.011	-0.019
Facebook Inc.	2.903	0.014	-0.011
Micron Technology Inc.	3.613	0.013	-0.017
Netflix Inc.	3.440	0.024	-0.011
Alibaba Group Holding Limited	2.758	0.011	-0.022
JP Morgan Chase & Co.	3.292	0.078	-0.028
Snap Inc.	4.932	0.017	-0.018
Alphabet Inc.	5.060	0.022	-0.025
Square Inc.	2.122	0.022	-0.011
Twitter Inc.	4.893	0.032	-0.011
Activision Blizzard Inc.	4.799	0.021	-0.037
Intel Corporation	2.802	0.042	-0.033
Gamestop Corporation	2.355	0.104	-0.281
Beyond Meat Inc.	6.400	0.012	-0.031
Palantir Technologies Inc. Class A	6.442	0.041	-0.074
Walmart Inc.	1.785	0.070	0.033
Virgin Galactic Holdings Inc.	8.481	0.016	-0.022
Morgan Stanley	1.854	0.069	0.029
Goldman Sachs Group Inc.	5.945	0.026	-0.028

Table 6: Boosted decision tree, next day returns, with WSB data.

Company	RMSE	R^2	adjusted R^2
Apple Inc.	2.029	0.369	0.367
Advanced Micro Devices Inc.	8.392	0.148	0.145
Tesla Inc.	4.064	0.146	0.142
Amazon.com Inc.	2.690	0.258	0.255
Facebook Inc.	3.610	0.214	0.211
Micron Technology Inc.	3.657	0.240	0.238
Netflix Inc.	4.766	0.130	0.126
Alibaba Group Holding Limited	2.792	0.205	0.201
JP Morgan Chase & Co.	2.861	0.469	0.467
Snap Inc.	7.102	0.029	0.021
Alphabet Inc.	6.880	0.085	0.078
Square Inc.	2.723	0.411	0.409
Twitter Inc.	4.495	0.242	0.237
Activision Blizzard Inc.	4.781	0.124	0.120
Intel Corporation	3.140	0.175	0.172
Gamestop Corporation	3.423	0.359	0.356
Beyond Meat Inc.	6.826	0.090	0.087
Palantir Technologies Inc. Class A	8.291	0.060	0.041
Walmart Inc.	2.15	0.184	0.181
Virgin Galactic Holdings Inc.	12.595	0.022	0.019
Morgan Stanley	2.093	0.174	0.170
Goldman Sachs Group Inc.	5.111	0.112	0.104

Table 7: Neural network, same day returns, without WSB data.

Company	RMSE	R^2	adjusted R^2
Apple Inc.	2.447	0.426	0.412
Advanced Micro Devices Inc.	8.806	0.171	0.152
Tesla Inc.	4.897	0.111	0.091
Amazon.com Inc.	2.336	0.353	0.334
Facebook Inc.	3.255	0.230	0.213
Micron Technology Inc.	3.056	0.193	0.175
Netflix Inc.	5.319	0.097	0.065
Alibaba Group Holding Limited	3.251	0.171	0.142
JP Morgan Chase & Co.	4.004	0.357	0.286
Snap Inc.	7.125	0.017	-0.015
Alphabet Inc.	7.502	0.074	0.034
Square Inc.	2.159	0.360	0.345
Twitter Inc.	4.171	0.202	0.175
Activision Blizzard Inc.	4.683	0.056	0.005
Intel Corporation	2.397	0.176	0.119
Gamestop Corporation	2.220	0.165	-0.192
Beyond Meat Inc.	6.623	0.084	0.044
Palantir Technologies Inc. Class A	6.833	0.107	-0.001
Walmart Inc.	2.135	0.146	0.114
Virgin Galactic Holdings Inc.	14.804	0.011	-0.027
Morgan Stanley	2.080	0.125	0.093
Goldman Sachs Group Inc.	4.918	0.091	0.046

Table 8: Neural network, same day returns, with WSB data.

Company	RMSE	R^2	adjusted R^2
Apple Inc.	2.838	0.044	0.040
Advanced Micro Devices Inc.	7.685	0.027	0.022
Tesla Inc.	5.050	0.017	0.013
Amazon.com Inc.	2.825	0.018	0.013
Facebook Inc.	2.769	0.018	0.015
Micron Technology Inc.	3.885	0.013	0.011
Netflix Inc.	5.886	0.007	0.002
Alibaba Group Holding Limited	3.149	0.011	0.006
JP Morgan Chase & Co.	2.430	0.023	0.019
Snap Inc.	7.056	0.025	0.017
Alphabet Inc.	6.090	0.013	0.005
Square Inc.	2.839	0.007	0.004
Twitter Inc.	5.178	0.015	0.011
Activision Blizzard Inc.	5.465	0.005	0.000
Intel Corporation	2.785	0.022	0.017
Gamestop Corporation	3.524	0.062	0.057
Beyond Meat Inc.	6.900	0.006	0.001
Palantir Technologies Inc. Class A	6.910	0.004	-0.016
Walmart Inc.	2.333	0.039	0.035
Virgin Galactic Holdings Inc.	11.712	0.005	0.000
Morgan Stanley	2.267	0.072	0.067
Goldman Sachs Group Inc.	5.022	0.047	0.035

Table 9: Neural network, next day returns, without WSB data.

Company	RMSE	R^2	adjusted R^2
Apple Inc.	2.509	0.017	-0.007
Advanced Micro Devices Inc.	5.573	0.012	-0.010
Tesla Inc.	3.956	0.012	-0.011
Amazon.com Inc.	2.997	0.032	0.003
Facebook Inc.	4.409	0.033	0.013
Micron Technology Inc.	3.387	0.014	-0.009
Netflix Inc.	5.129	0.019	-0.016
Alibaba Group Holding Limited	3.197	0.008	-0.025
JP Morgan Chase & Co.	5.460	0.092	-0.013
Snap Inc.	7.152	0.033	0.000
Alphabet Inc.	5.557	0.011	-0.027
Square Inc.	2.820	0.050	0.021
Twitter Inc.	5.974	0.052	0.015
Activision Blizzard Inc.	5.332	0.048	0.003
Intel Corporation	2.436	0.018	-0.060
Gamestop Corporation	2.341	0.074	-0.323
Beyond Meat Inc.	8.594	0.013	-0.029
Palantir Technologies Inc. Class A	6.418	0.034	-0.082
Walmart Inc.	2.422	0.055	0.019
Virgin Galactic Holdings Inc.	12.056	0.012	-0.025
Morgan Stanley	2.528	0.062	0.026
Goldman Sachs Group Inc.	6.403	0.028	-0.021

Table 10: Neural network, next day returns, with WSB data.

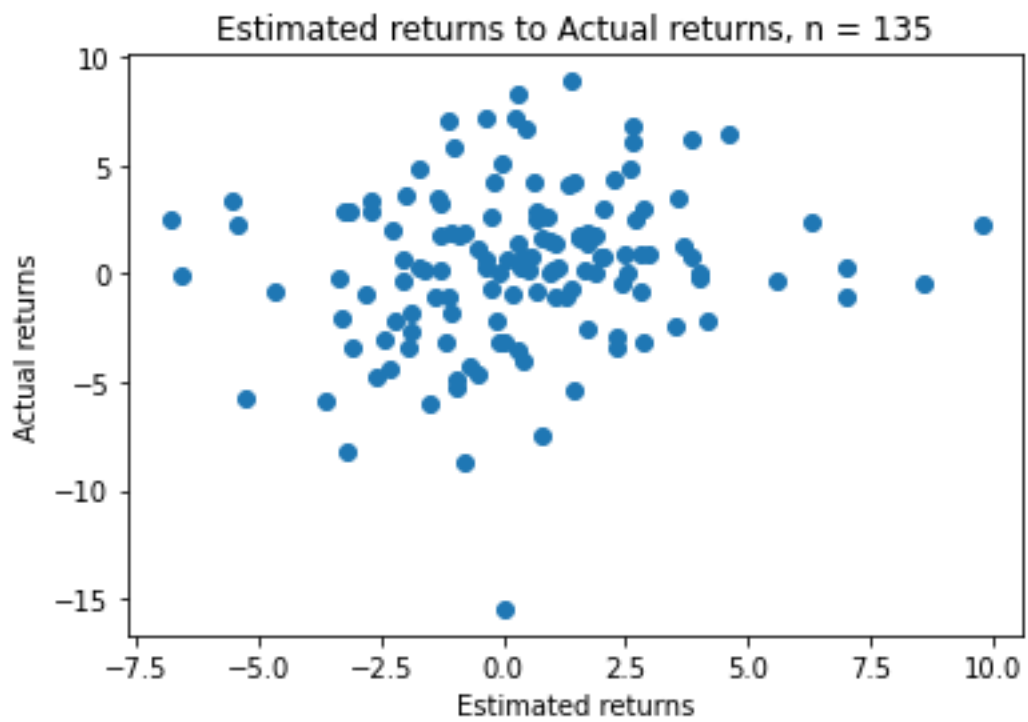


Figure 5: Scatter plot of Advanced Micro Devices, estimated returns to actual returns, using only reddit data (no index data). Same day. Correlation of 0.0255.

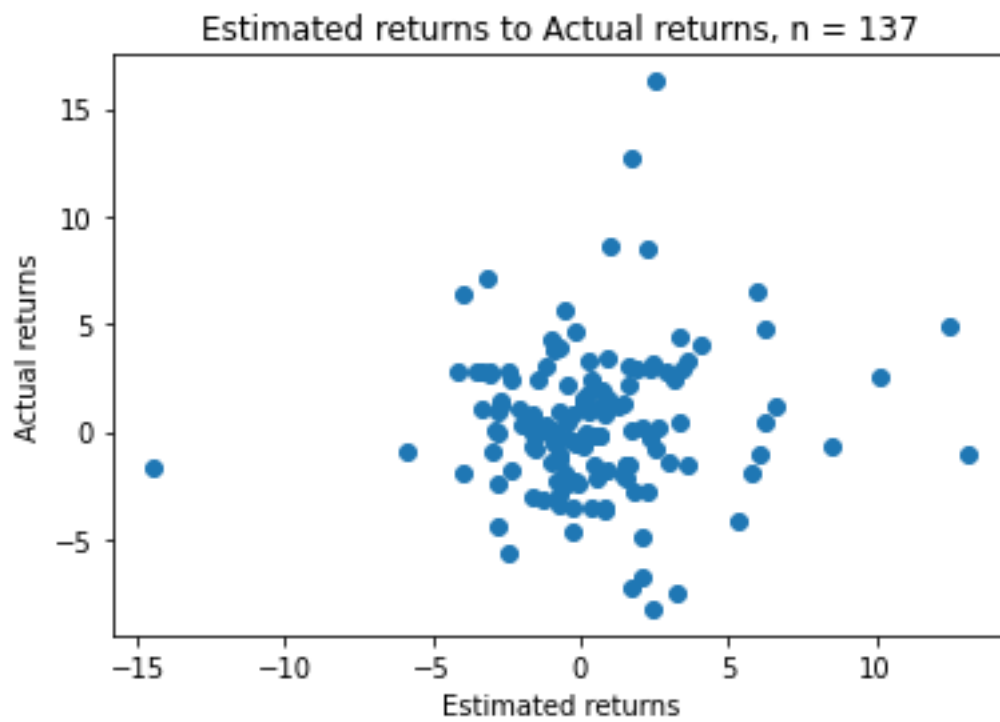


Figure 6: Scatter plot of Tesla, estimated returns to actual returns, using only reddit data (no index data). Same day. Correlation of 0.0411.

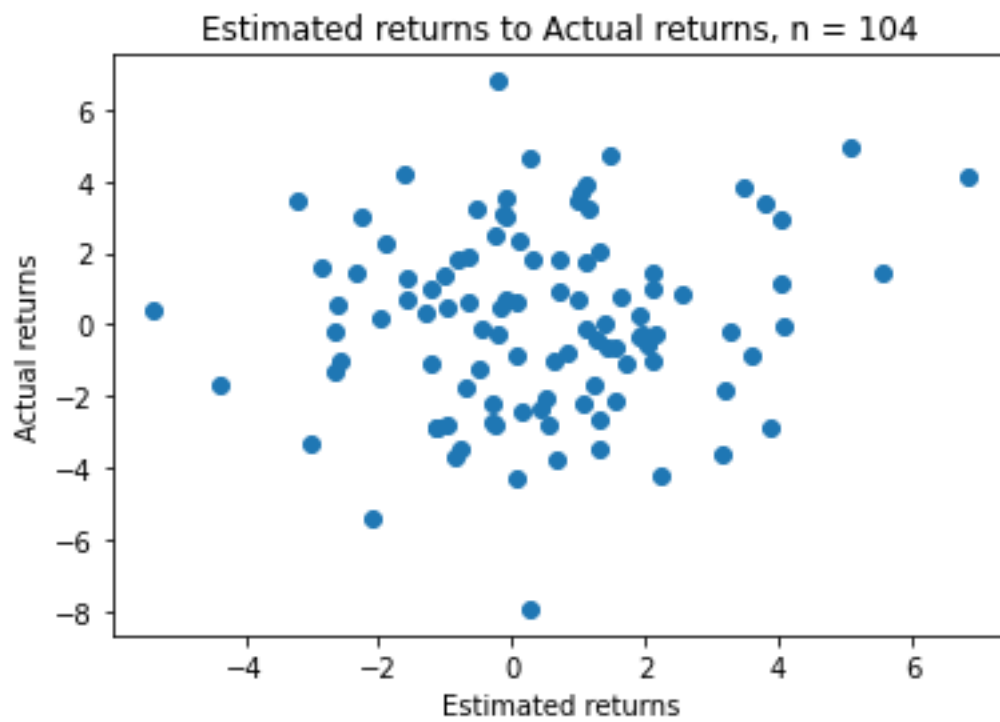


Figure 7: Scatter plot of Micron Technology, estimated returns to actual returns, using only reddit data (no index data). Same day. Correlation of 0.1148.

6 Discussions

6.1 Results

6.1.1 Empirical Asset Pricing Replication

As described in section 5.1, the predictive power of the replicated neural network on monthly data is high and the trading strategy by consistently taking long positions in the top ten stocks yielded a higher return than equal weighting between all stocks in the test. However, when trying the same with daily data, the prediction value is close to zero. This can mean two things; firstly, it might be that the paper and its findings are not replicable with daily data instead of monthly data. This might for example be due to larger spreads and variance in daily data, and small differences on a daily basis in many features which decrease the prediction power. Secondly, the neural network we implement is not appropriate or wrongly implemented. There are no recognized rules or models on how a neural network should be designed and constructed, which makes us question whether the paper is not replicable or if it is our neural network which is at fault. Nevertheless, since the prediction value is close to zero on the daily data, there is no reason for us to extend the data set used with the wallstreetbets features for the daily data. As for the monthly data, the timespan of wallstreetbets and the amount of activity of the forum only stretches a few years back, thus making it difficult to implement and extend a data set which goes back to 1980. This is primarily due to that the amount of data points would be small, especially considering wallstreetbets gaining a substantial size of data only in recent years.

6.1.2 Wallstreetbets Sentiment Analysis

We find that, on our analysis of same day returns and data with boosted decision trees, the RMSE decreases in most cases, with an average of 3.782 before the reddit data is added to 3.595 when we add the WSB data to our analysis. The R^2 generally increases with an average increase from 0.095 to 0.140 with the added reddit data. The R^2 has an arithmetic mean of 0.140 with the reddit data, which means that 14% of the data fit the regression model. The adjusted R^2 does not show a clear pattern of an increase or decrease. There is an average decrease from 0.089 to 0.086 when the reddit data is added. Gamestop has a significantly negative value of the adjusted R^2 with the reddit data (-0.168), which is likely due to the high variance in sentiment and activity of the comments and submissions. This high variance is consistent with Gamestop being a typical "meme-stock".

On the boosted decision tree analysis of next day returns, the RMSE generally decreases, going from an average of 4.245 to 4.017 when we include the reddit data. The R^2 shows an increase in most cases with an average of 0.011 to 0.031 with WSB. Overall, we can see that the RMSE is generally larger than in the same day analysis, and the R^2 is generally

smaller. The adjusted R^2 is close to zero without the reddit data, and in general becomes more negative when the WSB data is added. It decreases from a mean of 0.004 to -0.029. Gamestop once again stands out with a relatively large negative adjusted R^2 of -0.281. This model has an R^2 of 3.1%, which is significantly smaller than the 14% in the model without the reddit data.

In our neural network analysis of same day returns, we find that the RMSE does not show a clear pattern of an increase or decrease when the WSB data is added. The RMSE shows a slight increase when the WSB data is added, from 4.651 to 4.774. The R^2 is reduced with the WSB data, from 0.193 to 0.169, and the adjusted R^2 is reduced from 0.188 to 0.119. The R^2 indicates that 16.9% of our data fits our model.

The results of the neural network approach to the next day returns indicate that the RMSE increases on average from 4.754 to 4.848 when including the reddit data. The R^2 generally increases with the reddit data, from 0.023 to 0.033. The adjusted R^2 is often slightly positive without the reddit data, and turns negative when the WSB data is added, and goes from an average of 0.017 to -0.025. The R^2 here is 3.3%, indicating a lower explanatory value with the reddit data, than without.

The increased R^2 and decreased RMSE when adding the wallstreetbets data point to there being explanatory value in the WSB posts for the same day stock returns while using the boosted decision tree. We cannot establish a causality through our method though, since the gain is small. The explanatory value of the WSB posts for the same day stock returns could be interpreted in several ways. One could infer that the explanatory value stems from the WSB posts affecting the stock price of that day, for example. However, this seems unlikely since the people on WSB are retail investors and the timeframe stretches back to 2012, far earlier than when WSB could have had any impact on the stock market, due to its small size. Thus, we believe that the more likely option is that the posts are a reaction to the stock price of the day. For example, if a stock quickly drops or rises in price, it inspires members of the subreddit to create posts about that, creating a relationship between the forum posts and the stock returns of individual days.

For the next day-analysis through boosted decision trees, both with and without the reddit data, the RMSE is generally significantly larger than in the same-day analysis, and R^2 and adjusted R^2 are generally smaller than in the same-day analysis. This indicates that this model is performing worse when testing the predictive value for the next day, as compared to using the same day stock return. The adjusted R^2 has values close to zero without the reddit data, which points to it having practically no explanatory value. When we add the reddit data, the adjusted R^2 becomes notably negative, suggesting that adding the reddit data weakens the predictive power of the model even more. The conclusion that we draw from this is that you cannot successfully predict asset prices by

analyzing the posts on wallstreetbets through this method. This is of interest for those searching for means to gain a trading advantage through analyzing internet forum data. Yang, Zhu, Cheng (2020) find results that indicate a predictive value of internet posts for the stock trading volumes. Similar results are found by Sabherwal et al. (2011), who find that internet stock forum sentiment does have predictive value for stock trades. Also, Antweiler and Frank (2004) find that data from internet stock boards can predict market volatility. Antweiler and Frank (2002) even claim that it is possible to beat the market through analyzing data from Yahoo Finance, though it would be risky because of the high volatility. These papers differ from ours in method and type of data analyzed, and it would be of further interest to replicate our machine learning method and use more advanced sentiment algorithms for example, to further explore this area of research and perhaps get a different result.

The neural network gets higher a RMSE, and a lower R^2 and adjusted R^2 when adding the reddit data to the features whereas the boosted decision tree gets a higher RMSE and a higher R^2 . This might seem wrong and improbable but this is due to the neural network being far more fragile and sensitive to outliers, influential data points, and overfitting. Hence, the neural network can find patterns in the training data which do not exist in the complete data set and thus, fail in predicting the returns of the test data. This is likely due to the high volatility and skewness in the WSB data which the neural network gets lost in.

6.2 Methodology

Initially, we planned to add our analysis of wallstreetbets as a feature on the machine learning toolbox which replicates Gu, Kelly, and Xiu (2020). However, the replication of Gu, Kelly, and Xiu (2020) performs significantly worse while using daily data, than the monthly data. Since the paper cannot be replicated with the daily data, there is no value in adding an extension to an already faulty analysis. It would not generate any clear results as to the predictive value of wallstreetbets when the replication analysis is flawed from the start. Therefore, we decide to perform our analysis of wallstreetbets separately, in order to be able to eliminate as many possible erroneous variables as possible, and to distinctively decide the predictive power of wallstreetbets.

Our boosted decision trees prove more successful than the neural networks in our analysis of predicting asset pricing through analysis of WSB posts. This is surprising, since Gu, Kelly, and Xiu (2020) find that neural networks performed the best compared to other methods of empirical asset pricing, including boosted decision trees. There might be several reasons for this. Firstly, while the neural network is a more complex and sophisticated technique than the boosted decision tree, neural networks are to a larger extent subject to overfitting, which means that the model finds patterns or relation in the training data set

which do not exist in the test data set or the complete data set as a whole. This results in worse results when the neural network finds these patterns in the explanatory variables but these do not correspond to the same change in the response variable, which causes the larger error. Secondly, whereas the neural network is more sensitive to overfitting, the boosted decision tree is regarded as "robust" and can efficiently handle extreme outliers and influential data points, without the loss in accuracy. Our reddit data is probably of lower quality and subject to higher variance than the one used by Gu, Kelly, and Xiu (2020), explaining why the neural network performs worse. And lastly, a boosted decision tree is much easier to both calibrate and run, whereas a neural network is more like a "black box", where there are no general ways of efficiently choosing the amount of layers, nodes, and algorithms to yield the best result.

When determining the accuracy of the models, we are choosing to use RMSE, R^2 , and adjusted R^2 . This is due to several reasons, but mainly two. Firstly, these measures capture the accuracy of the model performance on the test data and not the complete data set. Secondly, they measure the qualities of the models which are interesting for us, accuracy/error, and explanatory value. Two examples of alternative ways of determining the explanatory value are correlation and mean absolute error, MAE. Correlation is a measure which tells us to what extent the linear least-square estimator of an explanatory and response variable explains the relationship between them. Thus, correlation is a great metric if there is, or is expected to be, a linear relationship between the variables. Otherwise, the correlation might be significant but it tells little about the actual distribution of the response variable and how well we are able to predict the actual outcome of a drawing from the distribution. However, it may be of interest to see the correlation between estimated values and actual values, but this should not be a primary measure to use. To summarize, correlation is a measurement not suitable to quantify the explanatory value of a model. MAE, in contrast to RMSE, quantifies the average of the errors but gives small and large errors equal weighting, whereas RMSE gives higher weighting to larger errors and smaller weightings to smaller errors. In this case, it is beneficial to penalize large errors since these would correspond to the days with the largest swings and possibility for profits and losses. Therefore, while MAE would fill the same function as RMSE and measure the mean error, RMSE is a more appropriate measure to use. Adjusted R^2 is the same as ordinary R^2 except it penalizes more features. When introducing new features to a model, it generally provides some level of explanatory value to the model at the cost of overfitting. The adjusted R^2 takes this into account which only increases when a new feature which introduces a substantial increase in explanatory value is introduced.

It is possible that some posts get associated with the wrong company, due to the algorithm picking up a name or ticker when it occurs in a post even though it is not the company which is the center of the post. A company name might get mentioned in a post whose meaning is not to describe the company, that is, the sentiment of the post is not intended

for that company. For example, both the words "Facebook" and "Twitter" have been mentioned frequently in posts not referring to the stocks of Facebook nor Twitter, but rather to them as a source of information. A more sophisticated algorithm could improve the matching of posts to the correct company and sentiment, for example by finding the subject of the sentence and matching it to different companies, or by some even more sophisticated sentence analysis tool. However, even though this would have been possible for this thesis, such sentence analysis tools are not tested, it would require a substantial amount of computational power, and lastly, be far more inconsistent than the current method. Therefore, we conclude it to be outside the scope of this thesis, but is an interesting topic for further research.

Our results are consistent with Tumarkin and Whitelaw (2001), since we find no explanatory value in using our stock forum data in order to predict stock returns later on, similar to their findings. By extension, our results do not coincide with Yang, Zhu, Cheng (2020), since they, unlike us, do find explanatory value in online stock board activity and trading volumes. We find that the posts which we analyze have some explanatory value due to the increased R^2 in the same-day analysis when the reddit data is added. This can be compared to the finding of Tumarkin and Whitelaw in that they find a correlation between abnormal activity on their forum and abnormal returns on the same day. Tumarkin and Whitelaw (2001) do not, however, investigate the causality of this. The authors discuss that a possible reason for this correlation may be that the forum reacts to the abnormal stock returns by abnormally high activity, and this scenario could be similar in our case. That is, the members of WSB may react to a change in stock price or another factor which is correlated with the price and decide to make posts about it, which then draw attention to that particular stock. This could explain the explanatory value of the posts and the same day stock returns. Further investigations into this to establish a causality would be an interesting extension.

The results which we obtain are in line with the Efficient market-hypothesis by Fama et al (1969), since they show no predictive value between the wallstreetbets posts analyzed by sentiment and frequency and the next day stock returns. The large decrease of the adjusted R^2 when adding the reddit data in the next day analysis, speak for that there is no usefulness in analyzing the WSB data in order to predict next day stock returns. By extension, it shows that it is not possible to gain a market advantage by analyzing this data through our method.

Since we find no influence of the WSB posts on the stock prices, as indicated by the lowered adjusted R^2 when including the reddit data on the next day stock returns, it contrasts with the notion of the Keynesian beauty contest (Keynes, 1936). If the Keynesian beauty contest holds true, we would have expected to see an effect of the postings on WSB on the stock returns the next day, since we would expect people to flock to the stocks which they

believe that others regard as popular and base their opinions on stocks' values on what they think others believe. For example, a high, positive activity on the forum around a certain stock might be interpreted as a sign of a positive market sentiment of that stock, and then inspire others to buy it as well. We would expect this scenario to lead to some predictive value of the WSB posts for the next day stock returns, and since we do not get any predictive value, it speaks against the idea of the Keynesian beauty contest in this context.

Weaknesses in our comparisons of our results with the Efficient market-hypothesis and the Keynesian beauty contest are for example the limited amount of data that we have used and the fact that the population of WSB is just a small fraction of all the world's stock buyers. Furthermore, people that are members of wallstreetbets also get influenced by other sources of stock information, such as other social media, newspapers, television, etc. This could lead to the posts of WSB actually having an effect on the population of WSB, but is too small to be picked up in this investigation. This effect would be in line with the Keynesian beauty contest, and in conflict with the Efficient market-hypothesis. Since the stock prices are influenced by an enormous amount of other factors, there might be too much noise to distinguish the effect of wallstreetbets' joint efforts by the means of this thesis.

One alternative to using the API could be to download all comments and submissions from <https://reddit.com/r/wallstreetbets> directly, while the downloaded data would be more multifaceted and accurate this way, it would only solve some of the problems while new would occur. When comparing the current method with using the API, there is, on average, a 40 per cent lower submission and comment count on reddit than what can be found with the API. This is mainly due to users deleting their comments and/or submissions and moderators deleting and cleaning up the subreddit. However, it is not possible to determine the moment when the content has been deleted and what the state of the thread was before it got deleted. For example, how many comments were in the thread, how many upvotes/downvotes had all the comments and submissions got, and how visible were they to different users? The content which can be accessed through reddit today is undeniably of higher data quality than what can be accessed through the API since it contains activity such as rating, comments, and upvotes/downvotes. The votes, the type of post (submission or comment), and the number of comments are examples of factors which likely affect how much attention a post gets, and by extension its potential of influencing sentiment. However, this approach would not have been scientifically appropriate to use since the thesis is about predicting stock movement and see whether wallstreetbets can help with predicting stock movements. Then it is not of interest to use historic data where 40 percent of data is missing, since there might be a structural relation between the submissions and comments that are deleted and those which are not. The method of accessing the data through the API is not optimal, but in this case preferable since

it catches all published comments and submissions on the forum and is thus far more consistent.

6.3 Future work

This thesis gives some insight of the relationship between wallstreetbets and the stock market, however the wallstreetbets data only stretches to the first of January 2021. Gamestop has experienced massive volatility during the first quarter of 2021 and it would be of interest to look into what happened there, and a more comprehensive and in-depth analysis of Gamestop in particular. It would be of interest to explore if it would have been possible to predict the increase in volatility and stock movement of January and February, and then use the same models and techniques to try to predict upcoming volatility. Hence, one interesting area of future work would be to look more specifically into certain "meme-stocks" and how the memes and activity have affected, or been affected, by the stock performance. However, this would require a more sophisticated way of determining the sentiment of the comments and submissions, as the Gamestop quickly transformed into a decentralized movement against the hedge funds Melvin capital and Citadel, making it difficult to determine and quantify the sentiment.

Additionally, the Hugging Face sentiment analysis transformer has been trained and developed on ordinary sentences and conversations. However, wallstreetbets is known for a vulgar language with frequent swearing and derogatory words. This creates a need for retraining of the transformer so it becomes specially developed to be able to determine the sentiment of the wallstreetbets comments and submissions. For example, the sentence "Shit is getting spring loaded to blast off to the fucking moon" is considered by wallstreetbets to be a positive sentence, indicating trust and belief in the stock, while most people would consider the sentence to be negative. DistilBERT labels this comment with a sentiment index of -0.995, that is, significantly negative. The solution to this would be to feed the model with labeled sentences which are characteristic to the dialogue on wallstreetbets. However, this labeling would have to be done manually which would be time consuming and realistically not done in a comprehensive way, which would lead to large fluctuations and volatility in the day-to-day average sentiment. Therefore, to train the sentiment transformer to more accurately determine the sentiment of wallstreetbets would be an area of high interest for further studies to determine and see whether the explanatory value would be even higher.

Lastly, it would be of interest to make a further segmentation of the time of comments and submissions in relation to the market. Now, we have segmented it per business day, but with further segmentation, it would be interesting to see how the sentiment and frequency relates to pre-market, aftermarket and during market hours, as well as how the aggregated sentiment of longer holidays relate to the market performance. This would require more

detailed stock and exchange data, as well as more data since the lack of comments and/or submissions referring to each stock would be much smaller and cause gaps in the data.

7 Conclusion

We explore the predictive value of adding sentiment and frequency data from popular companies on the internet stock forum wallstreetbets in addition to the S&P 500 index, to predict the stock returns of these popular companies through neural networks and boosted decision trees. We also replicate the empirical asset pricing by Gu, Kelly, and Xiu (2020) with the extension of using daily instead of monthly data.

The empirical evidence found through the boosted decision trees suggests that the posts of wallstreetbets have some explanatory value for the same day stock returns. This could be as a result of the posters reacting to the stock return by writing about it on the forum. The posts on wallstreetbets do not have any predictive value for predicting the stock movements of the next day. This is in line with the Efficient market-hypothesis. It furthermore reinforces the result of Tumarkin and Whitelaw (2001), in that it points to the fact that it is not possible to gain an advantage and beat the market by analyzing data from internet stock forums. At the same time, it contrasts with the Keynesian beauty contest (Keynes, 1936) as there is no found influence of the online posts that seem to affect other buyers of the stock market.

Unlike Gu, Kelly, and Xiu (2020), our boosted decision tree performs better than our neural networks. While there are many possible answers to why that is, the most likely is that the neural network is sensitive to overfitting and influential outliers, which decreases the predictive power when run on the test data. Additionally, *Empirical Asset Pricing via Machine Learning* is successfully replicated on monthly data but not on daily. This suggests that the paper either is not replicable with daily data due to noise such as larger variance in daily data which worsen the prediction, or that the neural network used is either unfitting to this analysis or implemented wrongly.

References

- Adamowski, Jan and Karapatak, Christina, 2010, Comparison of Multivariate Regression and Neural Networks for Peak Urban Water-Demand Forecasting: Evaluation of Different ANN Learning Algorithms, *Journal of Hydrologic Engineering*, 17(7), pp. 834-836.
- Anthony C. Ng, Zabihollah Rezaee, 2015, Business sustainability performance and cost of equity capital, *Journal of Corporate Finance*, 34, pp. 128-149.
- Antweiler, Werner, and Murray, Frank Z., 1993, Internet stock message boards and stock returns, Working Paper, University of British Columbia.
- Antweiler, Werner, and Murray, Frank Z., 2004, Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards., *The Journal of Finance*, 59(3), pp. 1259-1294.
- Brown, Abram, 2021, Founder Of WallStreetBets Discusses Why The Group Unleashed Chaos On GameStop—And Why He’s (Really) Exiled From Reddit, Published on [www.forbes.com](https://www.forbes.com/sites/abrambrown/2021/01/28/founder-of-wallstreetbets-discusses-why-the-group-unleashed-chaos-on-gamestop-and-why-hes-really-exiled-from-reddit/?sh=55927985211c). <https://www.forbes.com/sites/abrambrown/2021/01/28/founder-of-wallstreetbets-discusses-why-the-group-unleashed-chaos-on-gamestop-and-why-hes-really-exiled-from-reddit/?sh=55927985211c> (Retrieved 2021-05-01)
- Chai, T. and Draxler, R. R, 2014, Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature, *Geosci. Model Dev.*, 7, pp. 1247–1250.
- Fama, Eugene F., Fisher, Lawrence and Jensen, Michael C. and Roll, Richard W., 1969, The Adjustment of Stock Prices to New Information, *International Economic Review*, 10(1), pp. 1-21.
- Gelman, Andrew, Goodrich, Ben, Gabry, Jonah and Vehtari, Aki, 2019, R-squared for Bayesian Regression Models, *The American Statistician*, 73(3), pp. 307-309
- Gobler, Erin, 2021, What Is a Meme Stock?, Published on [www.thebalance.com](https://www.thebalance.com/what-is-a-meme-stock-5118074). <https://www.thebalance.com/what-is-a-meme-stock-5118074> (Retrieved 2021-05-01)
- Hayes, Adam, 2021, Retail Investor, Published on [www.investopedia.com](https://www.investopedia.com/terms/r/retailinvestor.asp). <https://www.investopedia.com/terms/r/retailinvestor.asp> (Retrieved 2021-05-01)
- Karim, Atm , Khaldoon, Albitar, and Elmarzouky, Mahmoud, 2021, A novel measure

of corporate carbon emission disclosure, the effect of capital expenditures and corporate governance, *Journal of Environmental Management*, 290.

Keynes, John Maynard, 1936, *The General Theory of Employment, Interest and Money*, Macmillan (London).

Lee, Isabelle, 2021, Published on Business Insider. <https://markets.businessinsider.com/news/stocks/melvin-capital-gabe-plotkin-took-home-846-million-in-2020-2021-2-1030067950> (Retrieved 2021-05-10)

Lintner, John., 1965, The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets, *The Review of Economics and Statistics*, 47(1), pp. 13-37.

Martin, Katie and Wigglesworth, Robin. 2021. Published on Financial Times. <https://www.ft.com/content/7a91e3ea-b9ec-4611-9a03-a8dd3b8bddb5> (Retrieved 2021-05-10)

Minka, Thomas P., 1999, Bayesian linear regression, *Proceedings - 2017 32nd Youth Academic Annual Conference of Chinese Association of Automation, YAC 2017*, pp. 376-381.

Mossin, Jan., 1966, Equilibrium in a Capital Asset Market. *Econometrica*, 34(4), pp. 768-783.

Peeters, Margot , Zondervan-Zwijnenburg, Mariëlle, Vink, Gerko and van de Schoot, Rens, 1999, How to handle missing data: A comparison of different approaches, *European Journal of Developmental Psychology*, 12(4), pp. 377-394.

Sabherwal, Sanjiv, Sarkar, Salil K., Zhang, Ying, 2011, Do Internet Stock Message Boards Influence Trading? Evidence from Heavily Discussed Stocks with No Fundamental News., *Journal of Business Finance & Accounting*, 38(9), pp. 1209–1237.

Sharpe, William F., 1964. Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk, *The Journal of Finance*, 19(3), pp. 425-442.

Shihao Gu, Bryan Kelly, Dacheng Xiu, 2020, Empirical asset pricing via machine learning, *The Review of Financial Studies*, 33, pp. 2223-2273.

Smith, Annabel, 2021, The Reddit revolt: GameStop and the impact of social media on institutional investors, Published on www.tradenews.com. <https://www.thetradenews.com/the->

reddit-revolt-gamestop-and-the-impact-of-social-media-on-institutional-investors/ (Retrieved 2021-05-10)

Suvojit Manna et al, 2017, A statistical approach to predict flight delay using gradient boosted decision tree, 2017 International Conference on Computational Intelligence in Data Science (ICCIDS), pp. 1-5

Tumarkin, Robert., Whitelaw, Robert F., 200,. News or Noise? Internet Postings and Stock Prices, Financial Analysts Journal, 57(3), pp. 41-51.

von Essen, Benjamin and Wu, Haohang, 2021, Predicting stock returns using LSTM, MFIN 330 Project, Stockholm School of Economics.

wallstreetbets, Published on www.reddit.com. <https://www.reddit.com/r/wallstreetbets/> (Retrieved 2021-05-10)

Yahoo Finance, Gamestop, 2021, published on www.yahoofinance.com.
<https://finance.yahoo.com/quote/GME/history?period1=1611187200&period2=1611964800&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true> (Retrieved 2021-05-10)

Yang, Xiaolan, Zhu, Yu., Cheng, Teng Yuan, 2020, How the individual investors took on big data: The effect of panic from the internet stock message boards on stock price crash, Pacific-Basin Finance Journal, 59.

8 Appendices

8.1 Appendix 1 - Extract of Corporate Dictionary

In table 11 is an extract of the dictionary we used to link names and tickers to the specific company. Since the appendix is large, only the companies which we analyzed in the end are included.

Word	Word Type	Company
AAPL	Ticker	Apple Inc.
apple	Name	Apple Inc.
AMD	Ticker	Advanced Micro Devices Inc.
Advanced Micro Devices	Name	Advanced Micro Devices Inc.
TSLA	Ticker	Tesla Inc.
Tesla	Name	Tesla Inc.
AMZN	Ticker	Amazon.com Inc.
Amazon	Name	Amazon.com Inc.
FB	Ticker	Facebook Inc.
Facebook	Name	Facebook Inc.
MU	Ticker	Micron Technology Inc.
Micron	Name	Micron Technology Inc.
NFLX	Ticker	Netflix Inc.
Netflix	Name	Netflix Inc.
BABA	Ticker	Alibaba Group Holding Limited
alibaba	Name	Alibaba Group Holding Limited
JPM	Ticker	JP Morgan Chase & Co.
JP Morgan	Name	JP Morgan Chase & Co.
SNAP	Ticker	Snap Inc.
Snap	Name	Snap Inc.
GOOGL	Ticker	Alphabet Inc.
alphabet	Name	Alphabet Inc.
SQ	Ticker	Square Inc.
Square Inc	Name	Square Inc.
TWTR	Ticker	Twitter Inc.
Twitter	Name	Twitter Inc.
ATVI	Ticker	Activision Blizzard Inc.
Blizzard	Name	Activision Blizzard Inc.
INTC	Ticker	Intel Corporation
Intel	Name	Intel Corporation
GME	Ticker	GameStop Corporation
gamestop	Name	GameStop Corporation
BYND	Ticker	Beyond Meat Inc.
Beyond Meat	Name	Beyond Meat Inc.
PLTR	Ticker	Palantir Technologies Inc.
Palantir	Name	Palantir Technologies Inc.
SPCE	Ticker	Virgin Galactic Holdings Inc.
Virgin Galactic	Name	Virgin Galactic Holdings Inc.
MS	Ticker	Morgan Stanley
Morgan Stanley	Name	Morgan Stanley
GS	Ticker	Goldman Sachs Group Inc.
Goldman Sachs	Name	Goldman Sachs Group Inc.

Table 11: Extract from the Dictionary with the companies which are analyzed.

8.2 Appendix 2 - Feature Examples

Variable	Explanation
1-month momentum	Momentum during previous month
Size	Market capitalization
12-month momentum	Momentum during previous 6 months
Change in 6-month momentum	Change in 6-month momentum
Maximum daily return	Higest daily return during that month
Industry momentum	Momentum within the stock's industry
Return volatility	Volatility of daily returns
Dollar trading volume	Total value of shares traded
Sales to price	Revenue over price of company
Share turnover	Number of shares bought and sold

Table 12: Top 10 important features according to Gu, Kelly, and Xiu (2020).